A Novel Trial Design for Studying Resilience to Clinical Stressors

Concurrent Natural History Cohorts for Calibration

Ravi Varadhan[1,2] Jiafeng Zhu[3] Karen Bandeen-Roche[2]

[1]Sidney Kimmel Comprehensive Cancer Center, Johns Hopkins Medicine
[2]Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health
[3]Department of Preventive Medicine, Northwestern University

March 03, 2026

Talk Overview

Goal: Propose a novel trial design for resilience research

1. **The Challenge:** Why standard approaches don't address our question
2. **The Solution:** Using natural history data to correct for bias
3. **The Evidence:** Validation that calibration works
4. **The Proposal:** A new design with concurrent natural history cohorts

Key Message

Prospectively planned natural history cohorts can serve as **calibration tools** — enabling efficient identification of **who recovers well** and **who doesn't**

Resilience Research: A Different Question

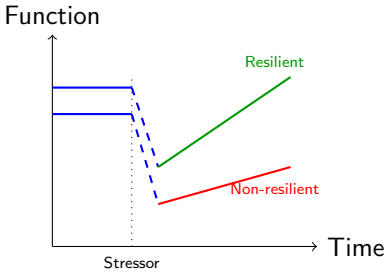
Resilience = Recovery after a major clinical stressor

NOT our question:
"Does TKR improve function on average?"
(Efficacy is well established)

OUR question:
Who recovers well and who doesn't?
→ *Predictors of resilience*

Why it matters:

- Personalized risk assessment
- Targeted prehabilitation
- Shared decision-making



Clinical stressors:

- Total knee replacement
- Cardiac surgery
- Hip fracture

The Analytical Challenge

Quantifying resilience: $\boxed{\text{Resilience} = Y_2 - Y_1}$

TKR Example:

- Y_1 = Function before surgery
- Y_2 = Function 1 year after
- $Y_2 - Y_1$ = Recovery = **Resilience**

Higher $Y_2 - Y_1$ = More resilient

Goal: Identify predictors of resilience

$$\underbrace{Y_2 - Y_1}_{\text{Resilience}} \sim Y_1 + \text{Age} + \text{Comorbidities}$$

Naive Analysis Fails

Mathematical coupling:
Y_1 on both sides $\rightarrow$ spurious correlation

Regression to the mean:
Extreme values regress toward average

Consequence:
Wrong predictors identified
(e.g., falsely implicates baseline)

Larger N doesn't help!

Our Solution: Correcting for Statistical Artifacts

Goal: Adjust the single-arm (surgery group) analysis to remove bias

The correction:

$$\text{True baseline effect} = \text{Naive} - \alpha_1$$
$$\text{True covariate effects} = \text{Naive} - \alpha_x$$

Where:

$$\alpha_1 = \frac{k\rho - m}{1 - R^2} + (m - 1)$$

$$\alpha_x = \frac{m - k\rho}{1 - R^2} \Sigma_{XX}^{-1} \Sigma_{1X}$$

From surgery group:

- ρ = correlation before & after
- R^2 = covariate prediction of baseline

Cannot estimate from surgery group:

- k = stability of function over time *without* surgery
- m = stability of covariate effects over time

Intuition

The correction subtracts what would have happened anyway (natural change) from observed associations

$\rightarrow$ What remains is the *true* treatment effect modification

Evidence: Naive vs. Corrected Results

Physical Component Summary (PCS) — 7,239 TKR patients

Naive Analysis:

Predictor	Effect	Sig?
Baseline function	-0.52	**Yes**
Age (per year)	-0.20	**Yes**
BMI (per unit)	-0.14	**Yes**
Male (vs Female)	$+0.47$	**Yes**
Comorbidity (≥ 2)	-3.19	**Yes**

Everything seems to matter!

Corrected Analysis ($k = m = 1$):

Predictor	Effect	Robust?
Baseline function	≈ 0	**No**
Age (per year)	-0.13	**Yes**
BMI (per unit)	≈ 0	**No**
Male (vs Female)	≈ 0	**No**
Comorbidity (≥ 2)	-1.52	**Yes**

Only age & comorbidity are robust

Key Finding

Baseline function is **not** a true predictor — its apparent effect is a statistical artifact

Non-Identifiability: Calibration with Natural History Data

Key insight: Bias arises because we don't know what *would have happened* w/o stressor

Problem:

- We observe: Y_1, Y_2 (pre/post stressor)
- We don't observe: Y_2^0 (outcome if *no* stressor)
- Correction requires k and m — but these depend on unobserved counterfactual (control information)

Solution:

- Use a **natural history cohort** (similar patients, no stressor)
- Learn how outcomes evolve *without* stressor
- Estimate k and m to enable correction

What natural history provides:

- Autocorrelation ρ_0 (stability of function over time)
- Variance ratio σ_{20}/σ_1 (variability change over time)
- Covariate-outcome relationship m

$\Rightarrow$ Allows estimation of k **and** m

Critical Distinction

Natural history cohort used for **calibration** (estimating k, m), not for direct comparison
$\Rightarrow$ Can be **smaller** than traditional control arm

Validation: TKR Registry + OAI Natural History Cohort

Test case: Can a natural history cohort enable valid analysis?

Data sources:

- **Stressor:** FORCE-TKR registry (N=5,148)
- **Natural history:** OAI cohort (N=592) — patients with OA, no surgery

Key findings:

- Calibrated estimates match gold-standard two-arm analysis
- **15–30% more precise** than traditional approaches

Predictors of resilience (corrected):

Factor	Predicts recovery?
Baseline function	No (spurious)
Age	**Yes** (older → less)
Comorbidity	**Yes** (more → less)

Naive analysis had falsely implicated baseline function!

A Novel Trial Design for Resilience Research

Stressor Cohort + Concurrent Natural History Cohort for Calibration

Traditional Two-Arm Design:

- Randomize/observe 1:1
- N = 500 stressor, N = 500 control
- Controls for efficacy comparison
- Designed for "Does it work?"

Mismatch for resilience research:

- We don't need to prove efficacy
- Large control arm is inefficient

Proposed Design:

- **Stressor cohort:** N = 500 (prospective)
- **Natural history cohort:** N = 300 (concurrent)
- Controls for *calibration only*
- Designed for "Who benefits?"

Advantages:

- 20% smaller total sample size
- Comparable precision for predictors
- Natural history only estimates k, m

Designing the Natural History Cohort

Purpose: Learn correction factors to enable valid analysis in stressor cohort

Eligibility:

- Same underlying condition
- Similar demographic range
- Candidates for stressor but don't receive it
- E.g., patients who decline, waitlist, conservative management

Sample size:

- Modest: N = 200–600 sufficient
- Only estimating a few parameters

Measurements (must align):

- Same outcome instruments
- Same timing (e.g., baseline, 1-year)
- Same key covariates

What we learn:

- How stable is function over time?
- How do age, comorbidity relate to natural change?

$\Rightarrow$ Correction factors for stressor analysis

Designing Your Trial – Checklist

When to use this design:

- Efficacy already established
- Primary goal is identifying predictors
- Can identify natural history patients

Natural history cohort:

- ☐ Similar eligibility criteria
- ☐ Same outcome measures
- ☐ Same assessment timing
- ☐ Key covariates collected
- ☐ N = 200–600 sufficient

Analysis steps:

1. Propensity weight if needed
2. Learn correction from natural history
3. Apply correction to stressor cohort
4. Report predictors with bootstrap CIs

Do NOT

Use naive change $\sim$ baseline regression — results will be biased regardless of sample size

Summary: A New Paradigm for Resilience Trials

1. **Resilience research asks:** Who recovers well? (not "Does it work?")
2. **Naive analysis fails:** Mathematical coupling and regression to the mean yield wrong predictors
3. **Our solution:** Use natural history cohort to estimate k and m for correction
4. **Proposed design:**
   - Stressor cohort + modest concurrent natural history cohort
   - Natural history for *calibration*, not comparison
   - 15–30% narrower CIs than traditional two-arm IPTW
5. **Validated:** TKR + OAI data confirms approach works

Why Calibration over IPTW?

IPTW integrates controls into outcome model $\rightarrow$ variance from both arms. Calibration uses controls only to fix $k, m \rightarrow$ all HTE inference from stressor arm alone.

R Package: `resilience` on CRAN: `https://CRAN.R-project.org/package=resilience`

Thank You!

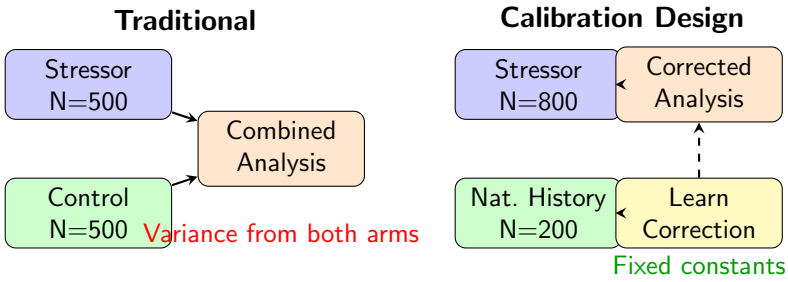
Questions?

ravi.varadhan@jhu.edu

References:

Varadhan, Zhu, Bandeen-Roche (2023) *Biostatistics*
Zhu, Xiang, Bandeen-Roche, Varadhan (2026) *Int. J. Biostatistics* (under revision)

Backup: Why Calibration Design is More Efficient



Why more efficient?

- Once correction is learned, it's *fixed* — no additional variance
- All inference from larger stressor cohort → more power for predictors
- Natural history only estimates a few parameters → modest N sufficient

Backup: Simulation Evidence

1,000 simulations comparing calibrated single-arm vs. two-arm approaches

	Two-Arm IPTW		Calibrated Single-Arm	
Predictor	RMSE	Coverage	RMSE	Coverage
Baseline effect	0.070	95.1%	**0.076**	95.2%
Age effect	0.621	95.5%	**0.174**	97.5%
BMI effect	0.716	95.9%	**0.530**	96.5%
Comorbidity effect	0.836	94.7%	**0.612**	97.0%

- Both methods: minimal bias, appropriate coverage
- Calibrated approach: **lower RMSE** (more precise estimates)